[[1]](#footnote-1)

Carga de Datos en RapidMiner y Hallazgos en una data set

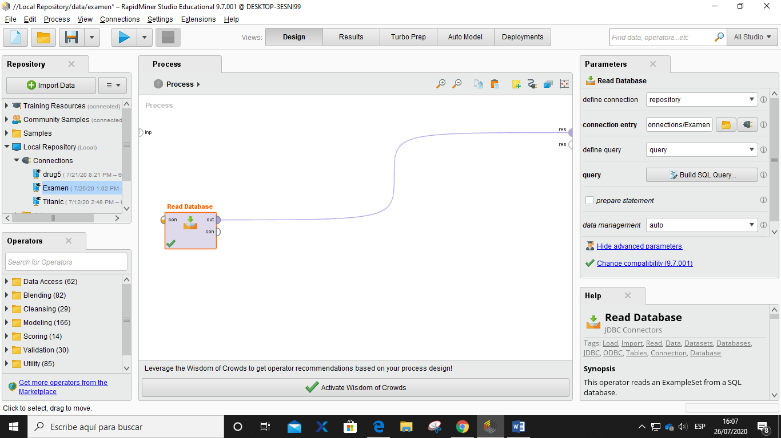
Eduardo Caiza-Andrés Proaño

*Resumen*—Básicamente lo que se va realizar es una breve explicación de cómo se realiza la carga de datos en la herramienta RapidMiner además de cómo crear diferentes modelos con el uso de la herramienta además se podrá ver como se realiza la predicción de algunos datos en función a otros datos de un mismo data set, además se podrá ver que tan preciso son las predicciones de los datos que producen la herramienta.

*Abstract*--Basically, what is going to be done is a brief explanation of how the data is loaded in the RapidMiner tool, as well as how to create different models with the use of the tool.In addition, it will be possible to see how the prediction of some data is performed based on others. data from the same data set, you can also see how accurate are the predictions of the data produced by the tool.

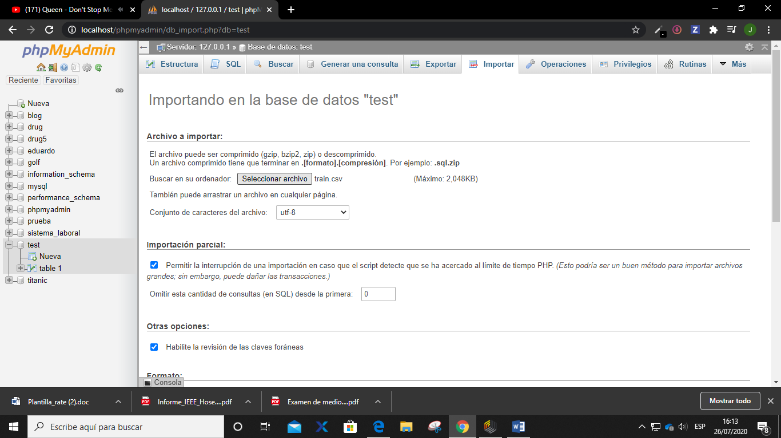
# Introducción

RapidMiner es una herramienta de Minería de Datos ampliamente usada y probada a nivel internacional en aplicaciones empresariales, de gobierno y academia. Implementa más de 500 técnicas de pre-procesamiento de datos, modelación predictiva y descriptiva, métodos de prueba de modelos, visualización de datos, etc.

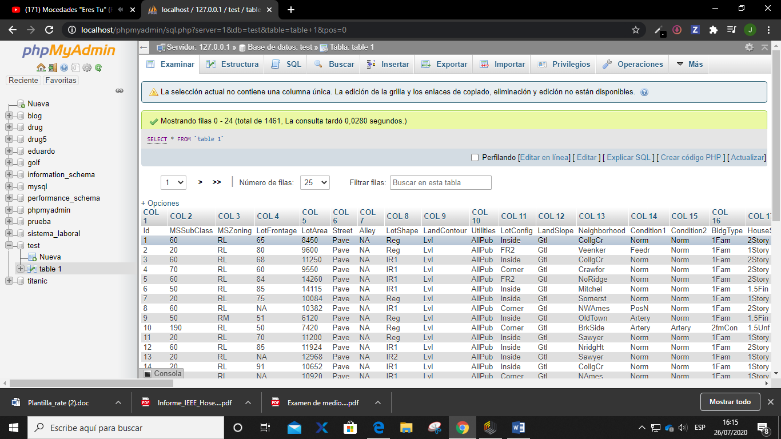


# Carga de datos en la herramienta desde mysql

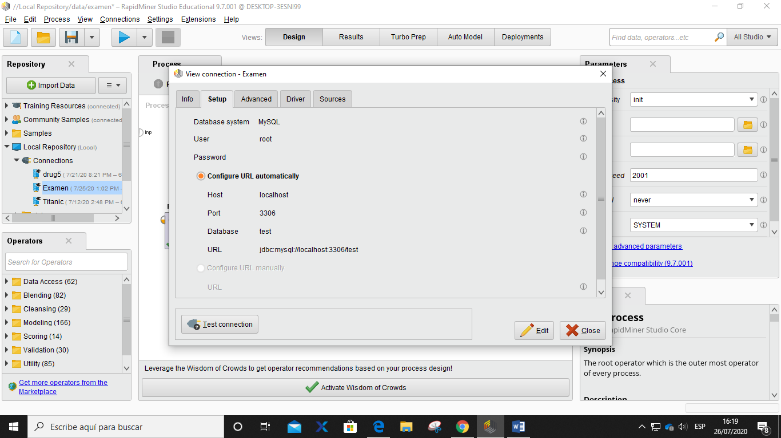
Para esto lo primero que vamos a realizar es crear una base de datos o importar una base de datos para probar la carga de datos se va a usar un archivo csv llamado train el cual va a ser cargado en MYSQL para esto vamos a crear una nueva base de datos llamado test en esta base de datos se procederá a importar el archivo csv



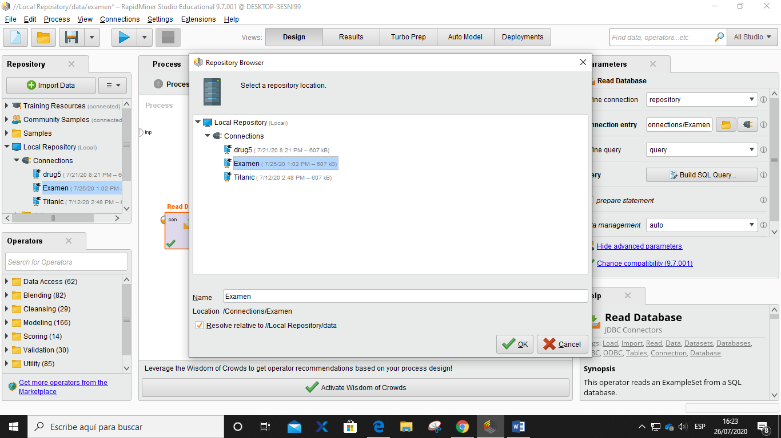
Una vez realziado esto podemos observar los diferentes datos que se crean en la base de datos en este caso estos datos seran insertados en una tabla llamada table 1 como se puede ver en la siguiente imagen



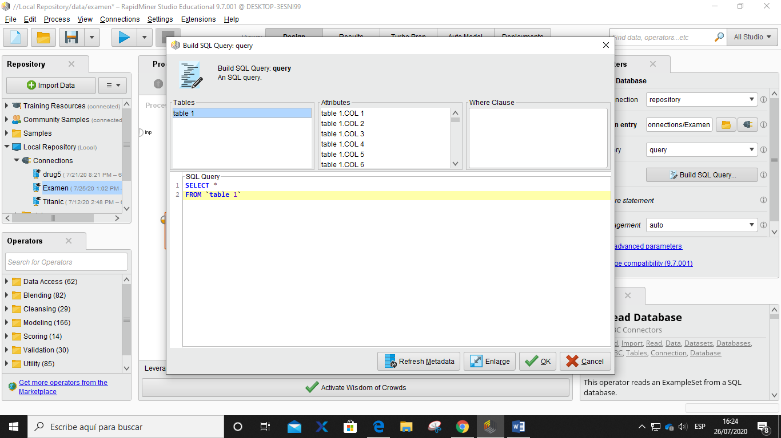
A continuación, lo que vamos a realizar es la creación de una conexión MYSQL en la herramienta como podemos ver a continuación



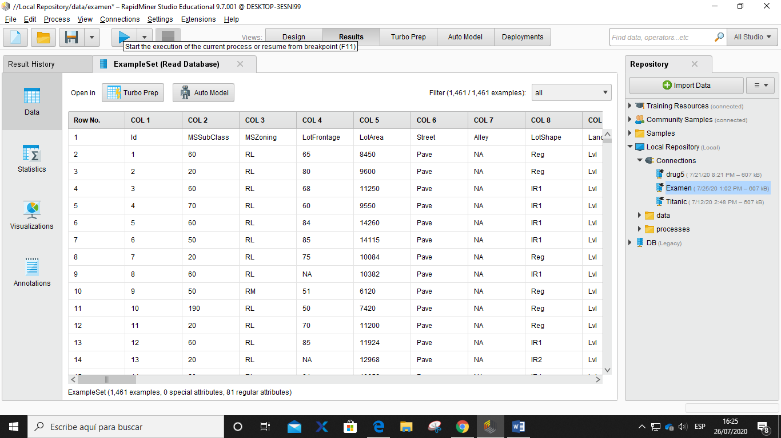
A continuacion lo que vamos a realizar es usar operador en este caso sera un read database a continuacion lo que vamos a realizar es las diferentes configuraciones para que nosotros podamos leer los diferentes datos de la base creada en MYSQL lo primero que vamos a realizar es seleccionar la conexion



A continuacion lo que vamos a hacer es seleccionar los datos que se van a mostrar en la herramienta en este caso seran todos



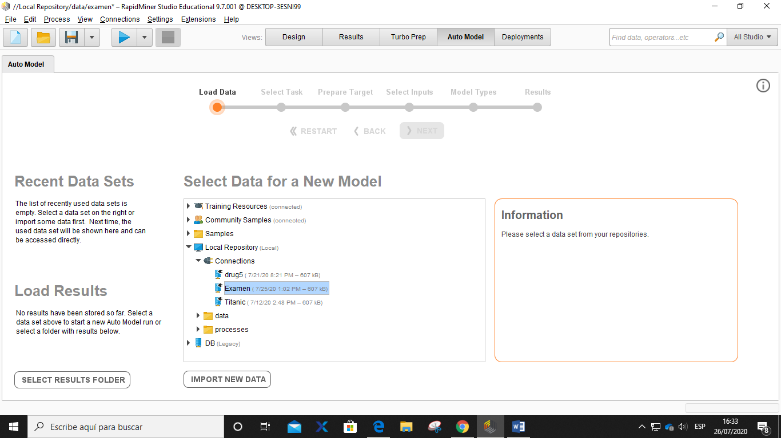
A continuacion lo que debemos hacer es probar si se cargaron los datos de MYSQL en nuestra herramienta.



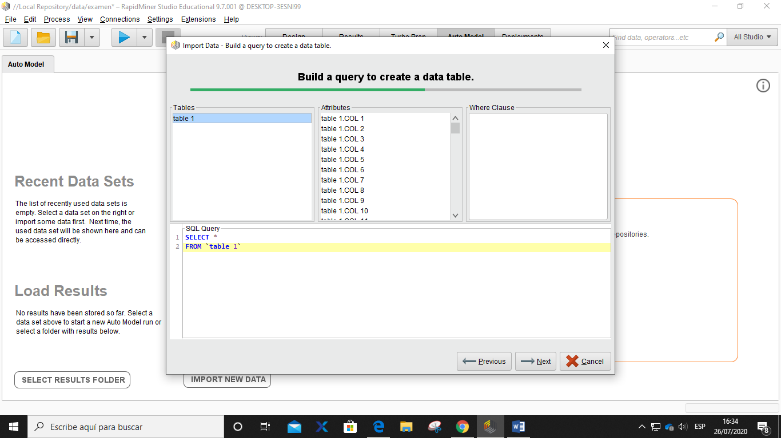
Con esto podemos ver que la carga de datos en nuestra herramienta fue exitoso ya que se muestran todos los datos de la base creada en MYSQL.

# CREACION MODELO

Lo primero que vamos a realizar es crear un nuevo proyecto con la opción auto model una vez seleccionado eso vamos a seleccionar la conexión que creamos

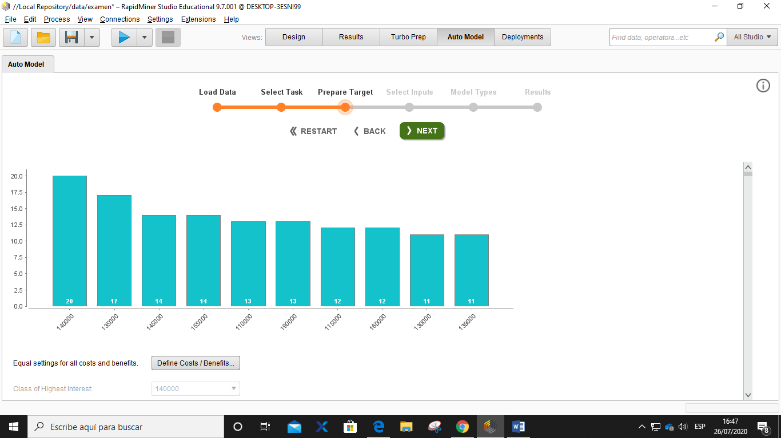


A continuacion lo que vamos a realizar es seleccionar la base tabla que contiene todos los elementos de la base de datos antes creada

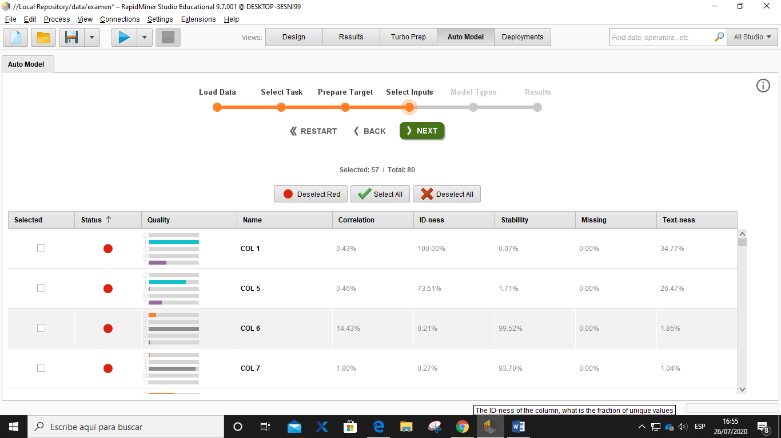


A continuacion lo que vamos a hacer es guardarla en data con el nombre de dataexamen una vez hecho esto lo que sigue es seleccionar finish una vez que termine se podra ver los diferentes datos de la base de datos a continuación lo que vamos a hacer es seleccionar la opción predicciones a continuación de esto seleccionamos la tabla que queremos predecir en este caso será SALE PRICE ***porque se escogió esta tabla básicamente se escogió porque la problemática del problema se basa es esta columna ya que es el precio de venta es lo que se calcula en base a todos los diferentes datos que están en otras columnas.***

A continuación, damos en siguiente y vemos un diagrama de barras



A continuacion nos sale un pantalla la cual nos permite escoger las diferentes columnas que se usaran para realizar la prediccion de datos en este caso se usaran todas las columnas que tienen datos numéricos ya que en base a esto se calcula el precio de venta por lo que en este caso no se tomara en cuenta las columnas que tengan datos como por ejemplo el nombre



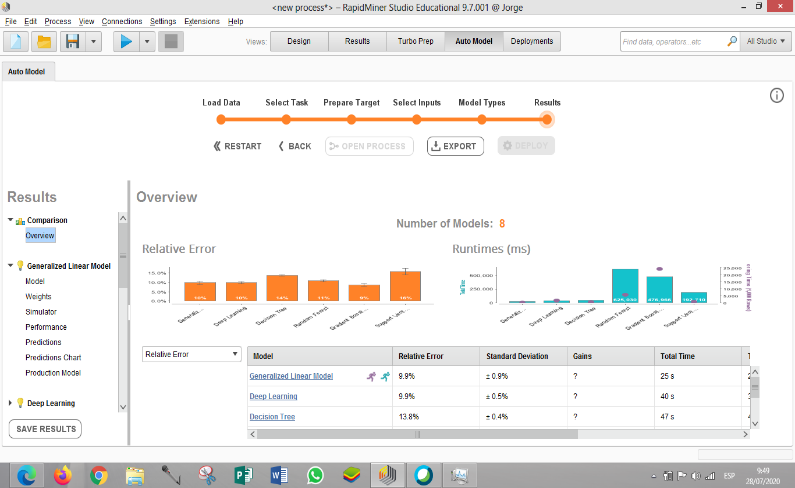
A continuacion lo que vamoa a hacer es seleccionar los algoritmos o modelos que se van a usar para realizar la predicción de los datos en este caso se usara los que están por defecto

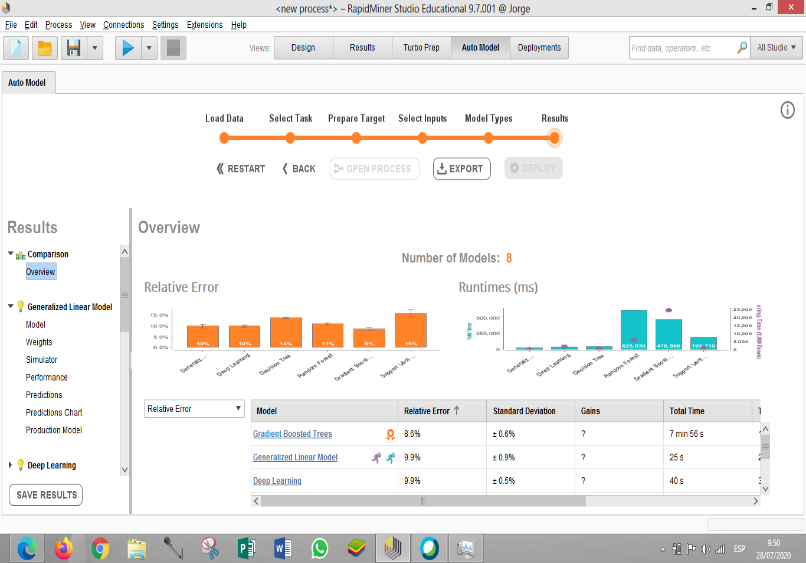
A continuación, lo que se va a realizar es crear la comparación del uso de diferentes modelos en base a esto podemos darnos cuenta de que la precisión de predecir dependerá de cada modelo por lo que se para esto se procederá a escoger tres modelos

El primero modelo que se escogió es el Deep Learning el cual nos da un error de 9.9%

El segundo modelo que se escogió es el Linear Model con un error de 9.9%

El tercer modelo que se escogió es el Gradient Boosted Tress con un error de 8.6%

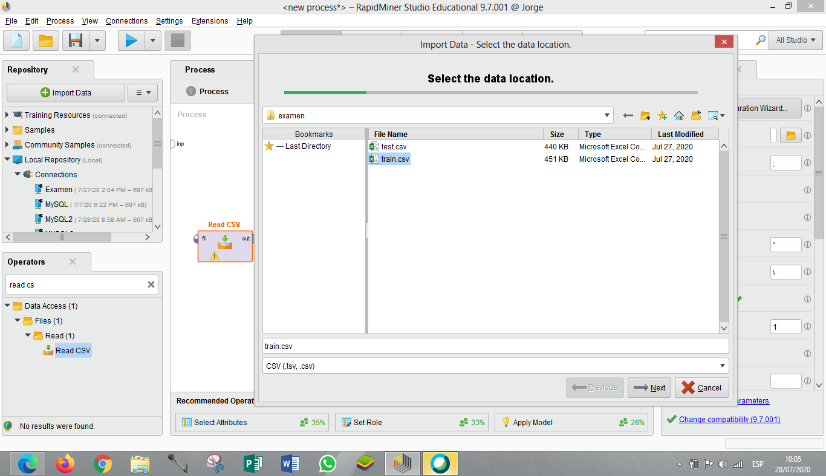




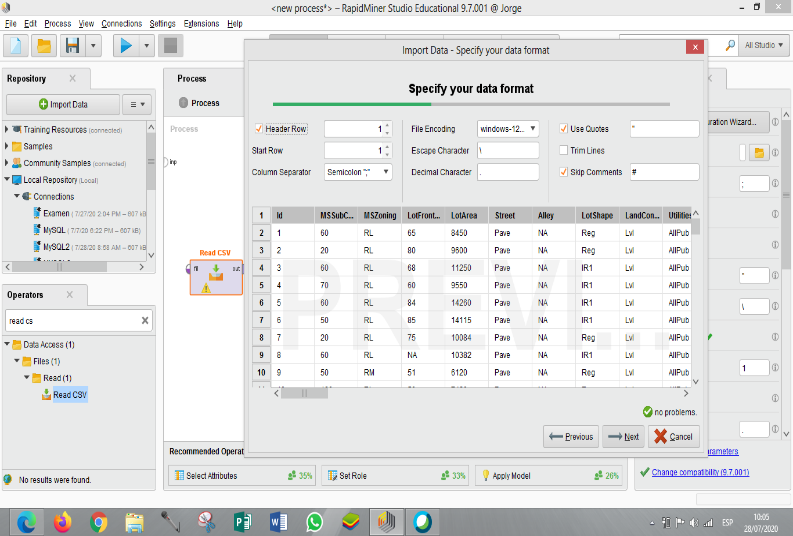
En las imágenes anteriores se puede ver el error producido por cada modelo por lo que al realizar una comparación entre los tres podemos concluir que el mejor modelo es el Gradient Boosted Trees ya que tiene un porcentaje de error menor la razón por la que se escogió estos modelos es que decidimos comparar un modelo con menor, mayor y otro modelo con un valor intermedio

# TRAINING

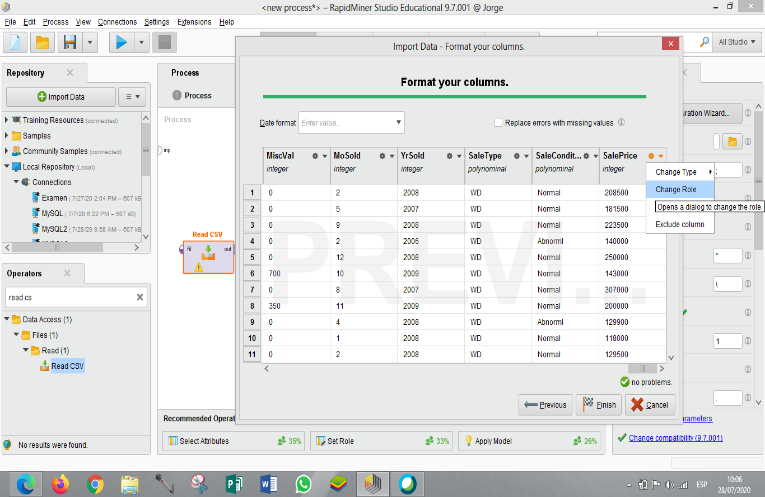
Para esto escogemos el archivo train



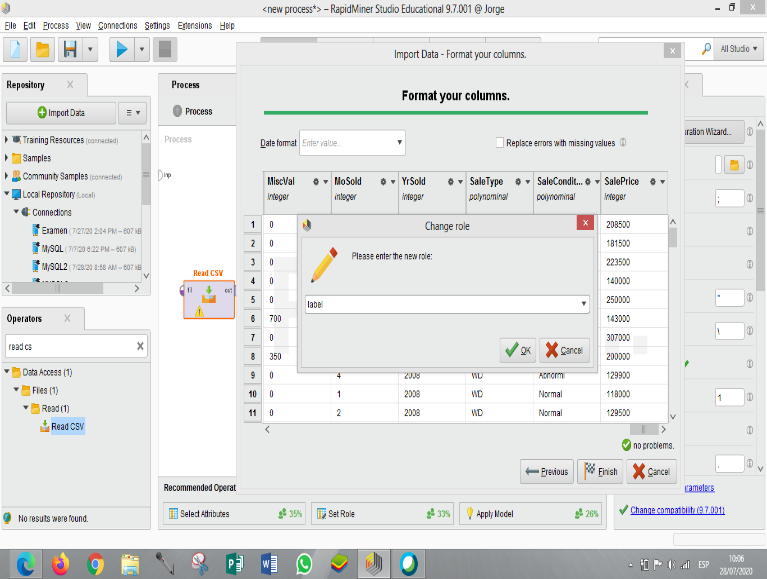
Aquí se puede observar los diferentes datos que contiene el train



Aquí se procederá a crear un campo label



En este caso será el campo Sale Price

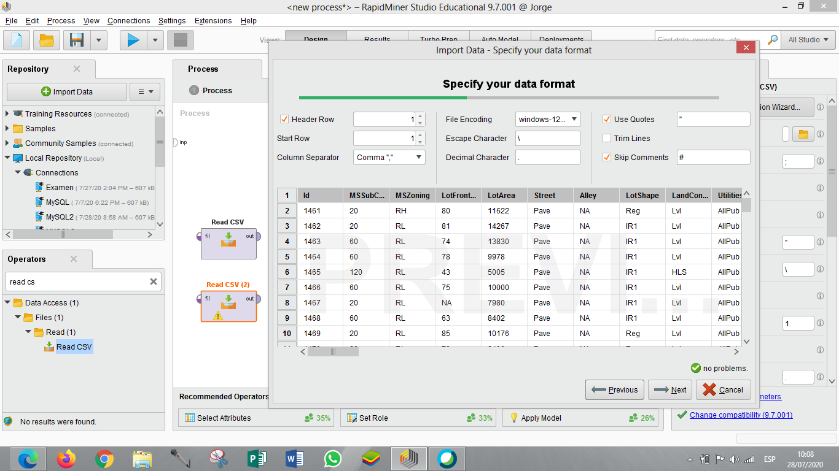


# TESTING

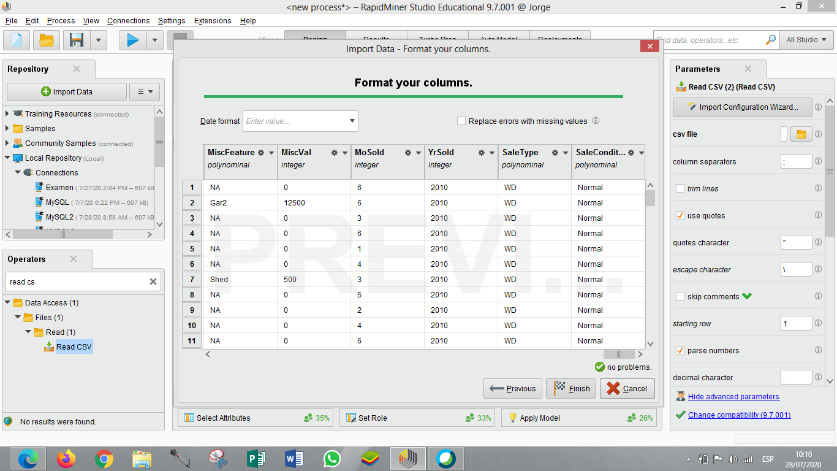
Lo primero que se realiza es seleccionar el archivo



Aquí se puede observar los diferentes datos que contiene el test

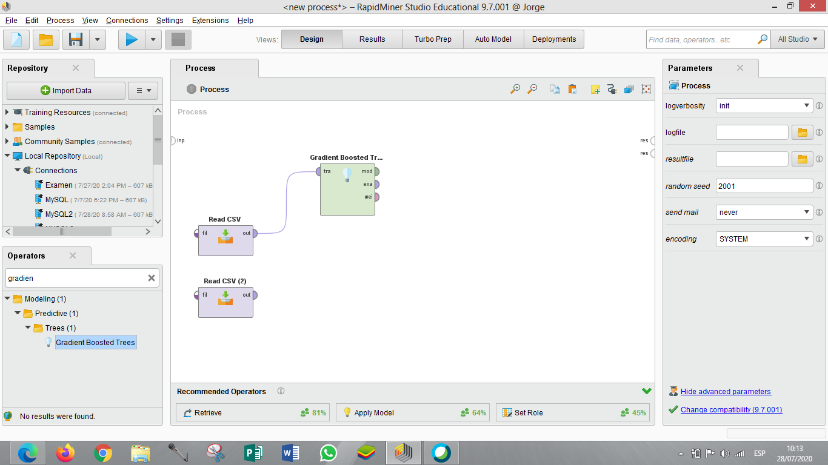


Aquí se procederá a crear un campo label

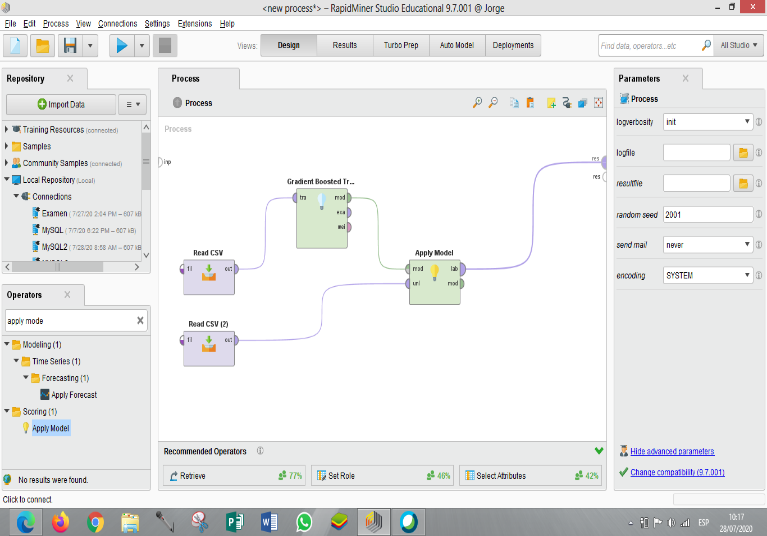


# ALGORITMO

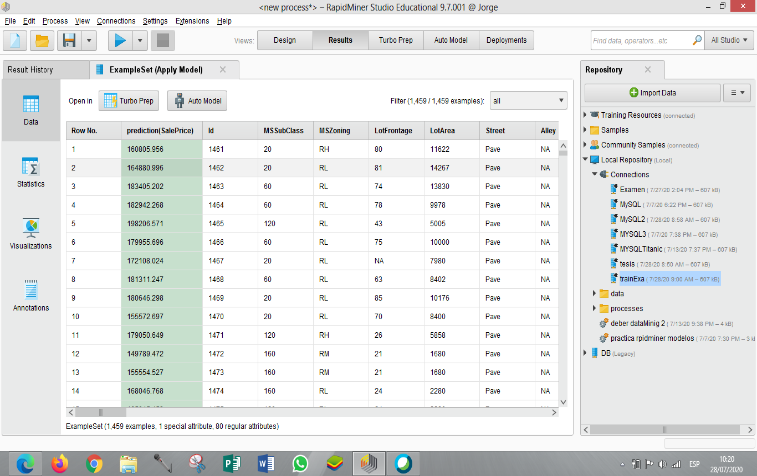
A continuación, se puede observar el algoritmo que se va a usar para aplicar los modelos anteriormente seleccionados



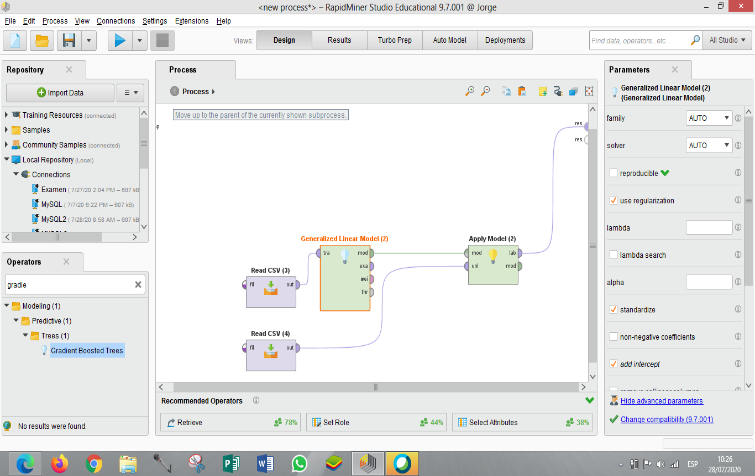
A continuación, tenemos el modelo que se creó para aplicar el modelo en este caso el Gradient Boosted Trees



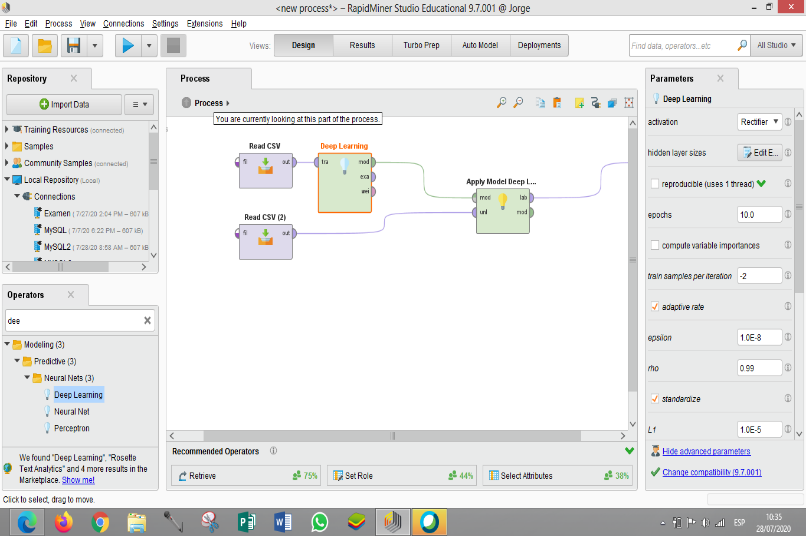
A continuación, procedemos a correr el proceso y podemos ver los datos que nos dan como resultado al aplicar este modelo



A continuación, se procederá a usar el mismo algoritmo para los otros modelos solo que para cada modelo se escogerá su correspondiente componente

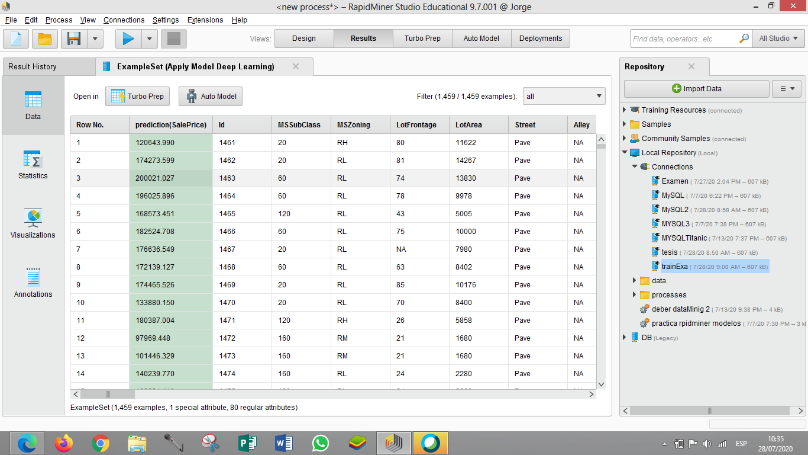


Linear Model

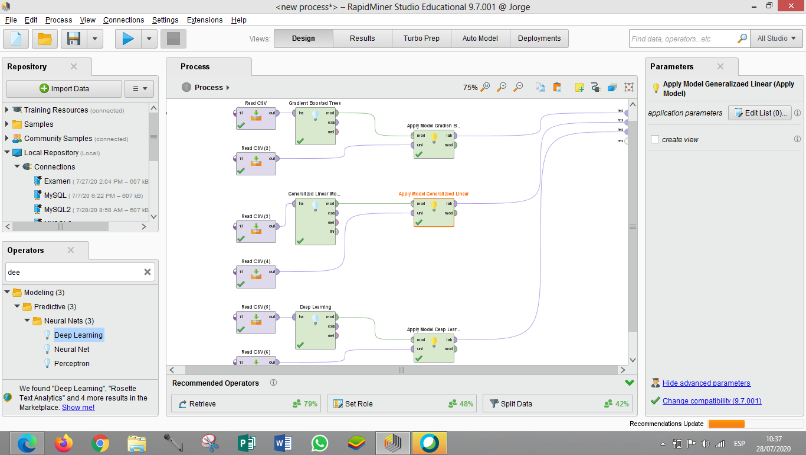


Deep Learning

A continuación, procedemos a correr el proceso y podemos ver los datos que nos dan como resultado al aplicar este modelo



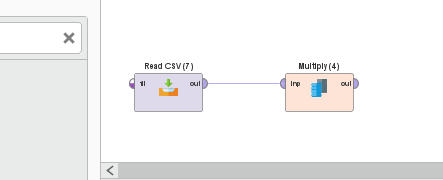
Por ultimo tenemos los tres modelos usando el mismo algoritmo solo que con un componente diferente correspondiente a cada modelo.



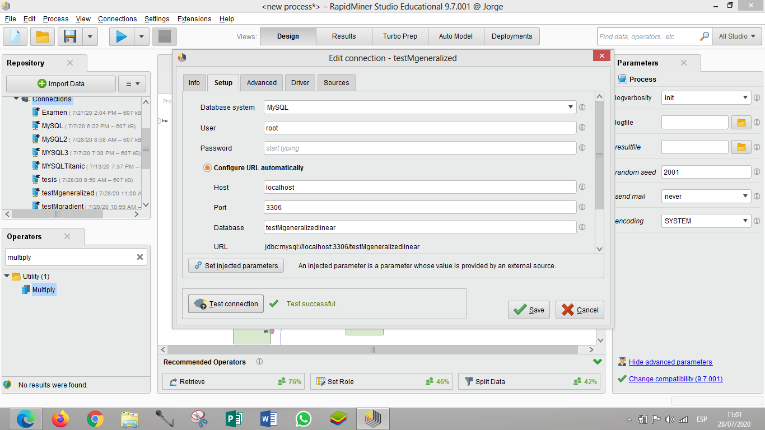
# PASAR LOS DIFERENTES DATOS A MYSQL Y MONGO DB

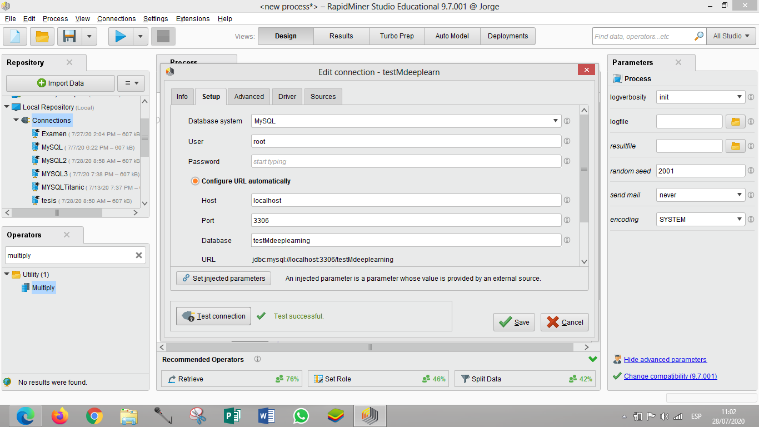
Lo primero que vamos a realizar es la creación de las diferentes conexiones para cada uno de los modelos que nos permitirá hacer la conexión a nuestra base de datos MYSQL y MONGODB

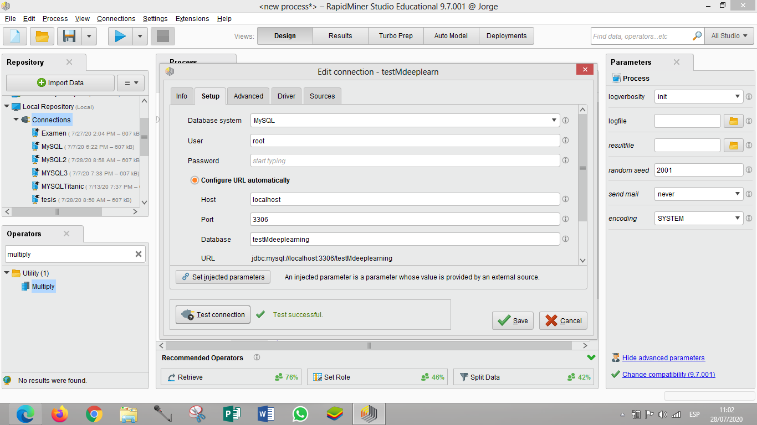
Lo primero que vamos a crear es el modelo para poder subir nuestros datos a mysql mediante el uso de del siguiente modelo



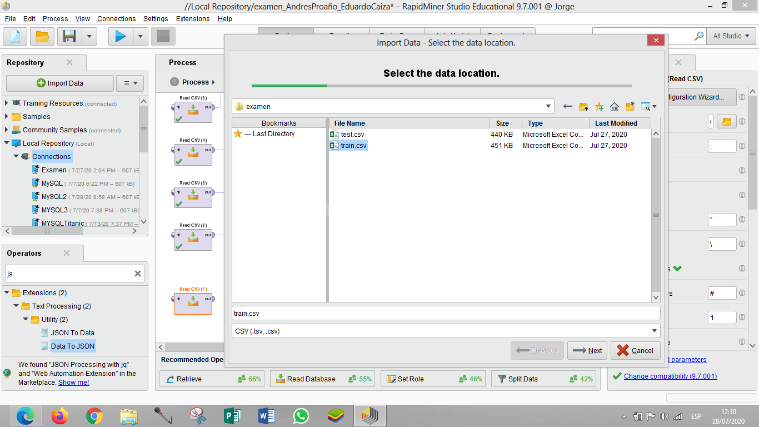
A continuación, realizamos las configuraciones para realizar la conexión





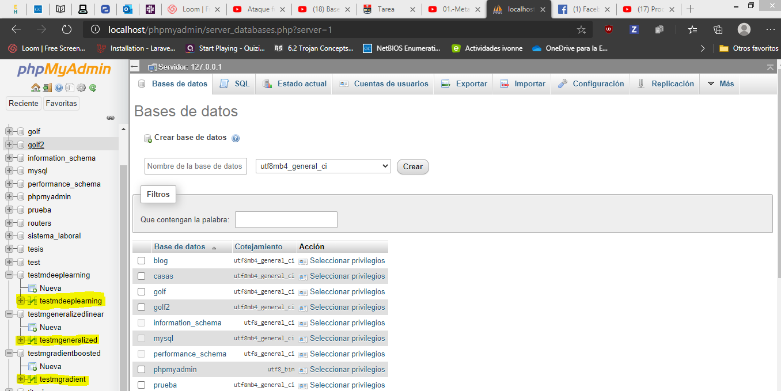


A continuación, seleccionamos los diferentes archivos que vamos a subir a nuestra base en MYSQL

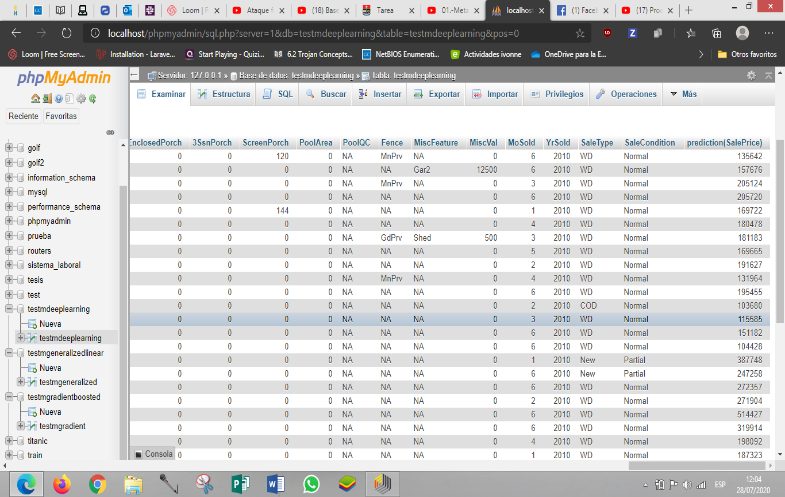


A continuación de esto seleccionamos la conexión y gracias al modelo podemos subir los diferentes datos a nuestra base como se puede ver a continuación

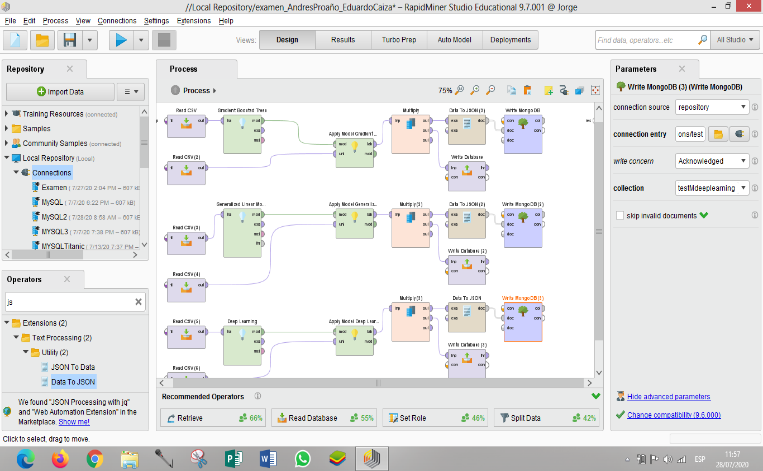
Aquí también se puede observar que se crearon bases de datos con el nombre correspondiente a cada modelo



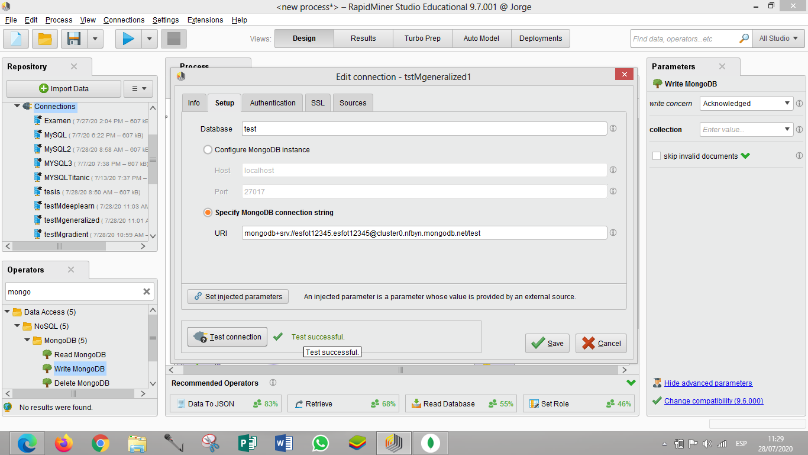
A continuación, tenemos los diferentes datos correspondientes a cada una de las bases de datos



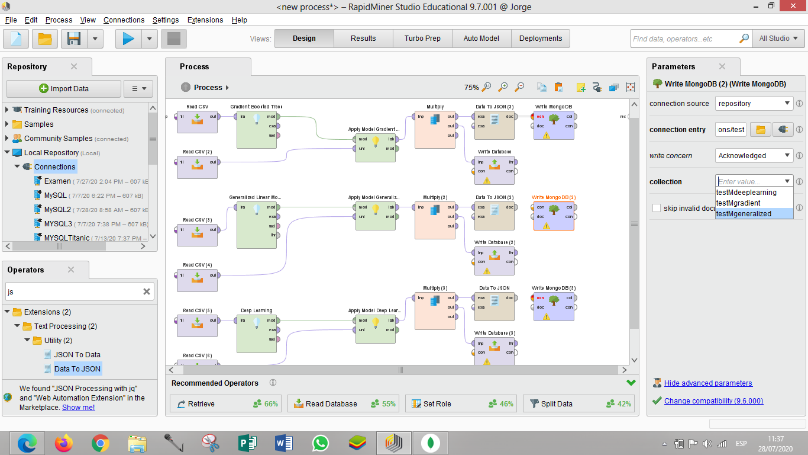
***A continuación, lo que vamos a crear es el modelo que nos permitirá subir nuestros datos tanto del test con del train a nuestra base de datos MONGODB***



A continuación, tenemos las diferentes conexiones creadas para la base de datos MONGODB esto se va a realizar para cada modelo

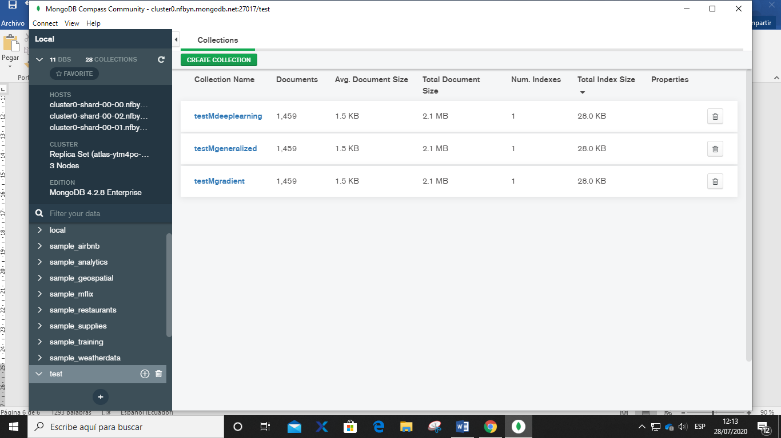


A continuación, podemos seleccionar la colección que nosotros queramos para poder subir nuestros datos

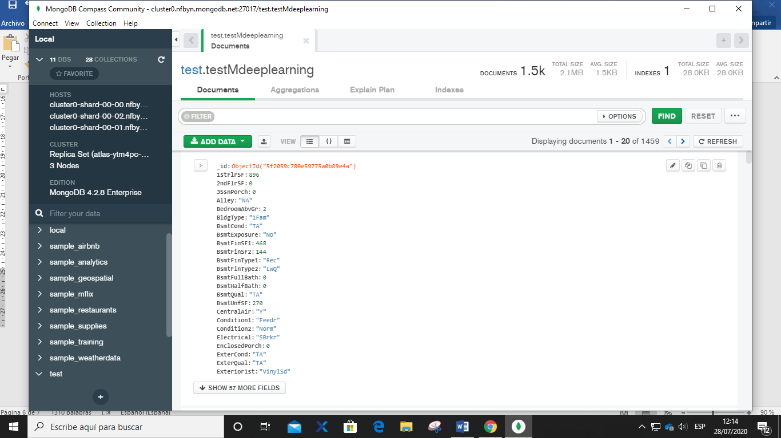


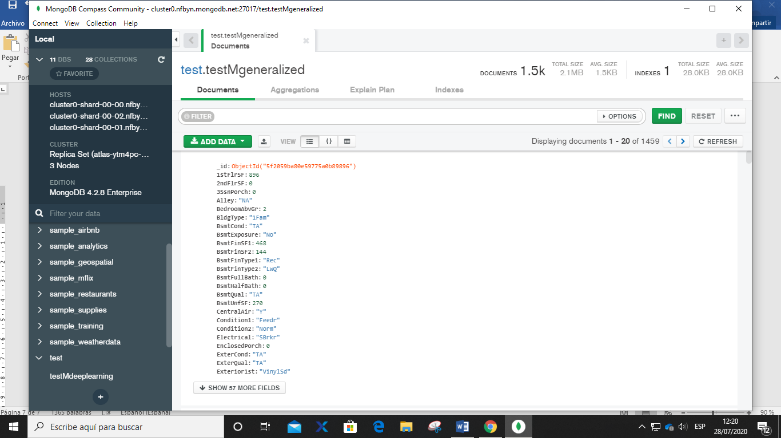
A continuación, podemos ver que los diferentes datos se subieron a la base de datos llamada test y además se subió los datos de cada uno de los modelos a su correspondiente colección

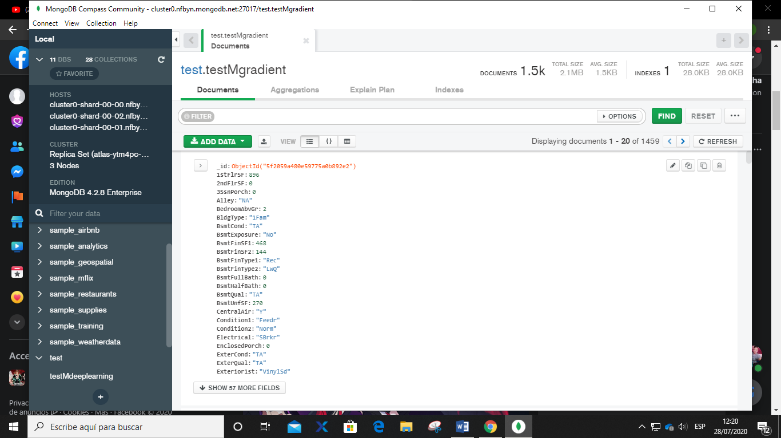
Aquí se puede observar la base de datos llamada test con tres colecciones correspondientes a cada modelo



A continuacion tenemos los diferentes datos cargados a cada colección ademas de esto se puede observar que los archivos se suben en un formato json esto se logro gracias al modelo antes creado el cual nos permite crear hacer salida de datos en el formato json el cual es adminitido en esta base de datos.

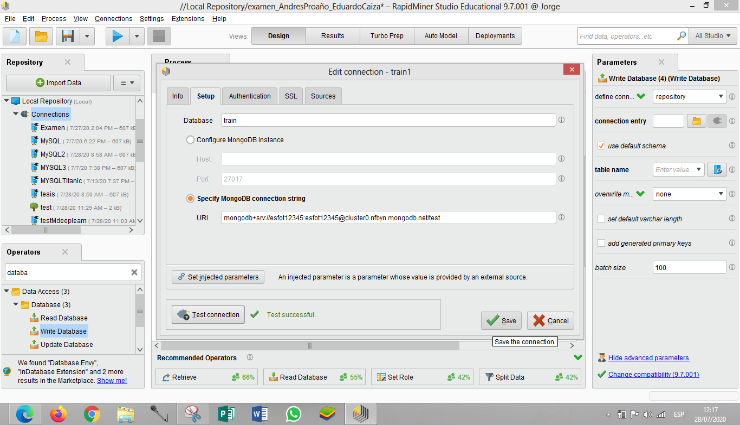




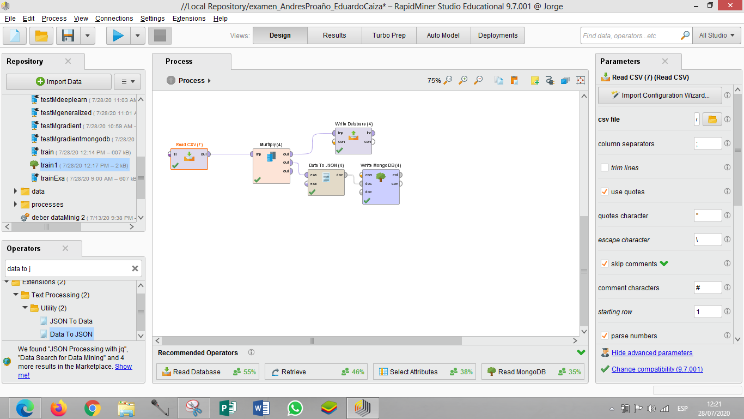


A continuación, vamos a realizar lo mismo pero estos datos van a ser subidos a una base de datos llamado train ya que es resultado de un testing

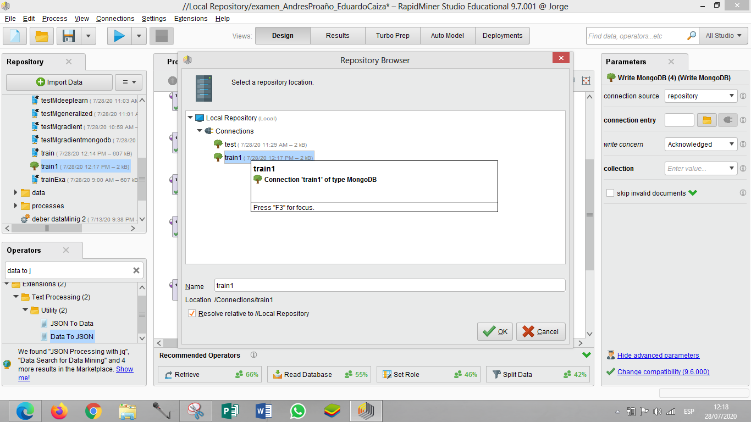
Lo primero es crear la conexión

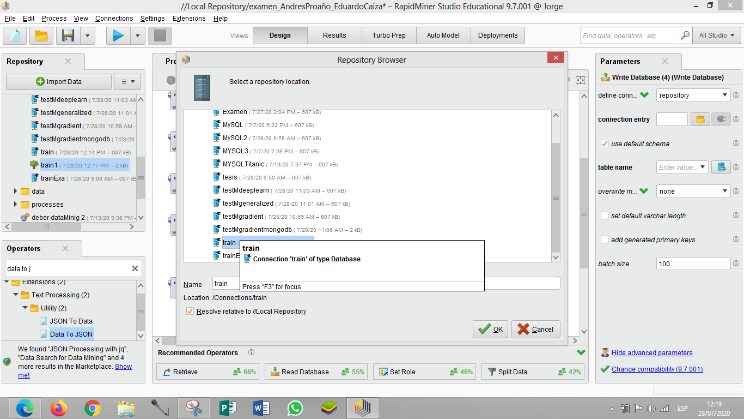


Creamos el modelo para realizar la operación

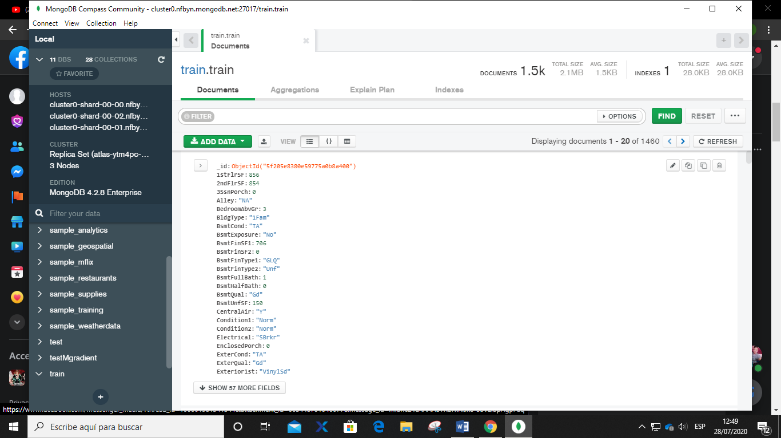


Seleccionamos la colección y la base de datos

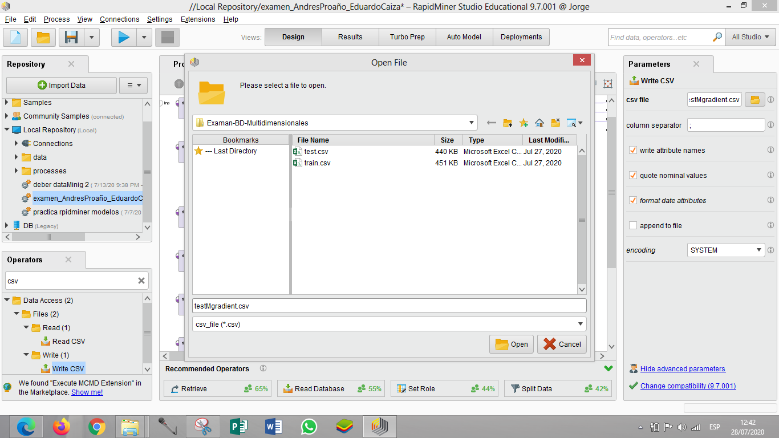




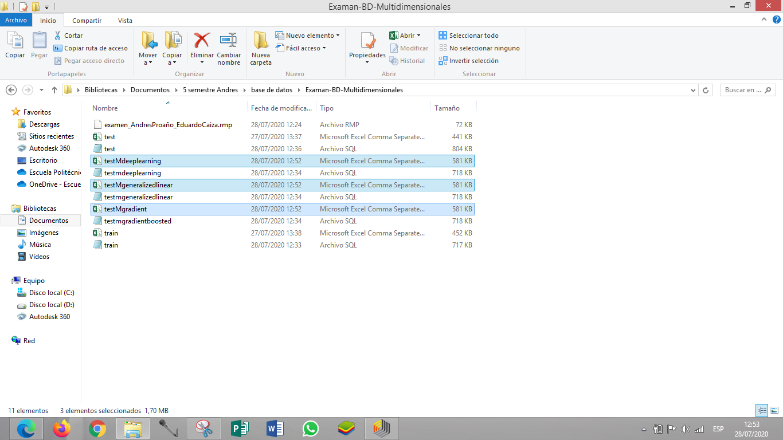
A continuación, tenemos los diferentes datos que se subieron a la base de datos llamada train



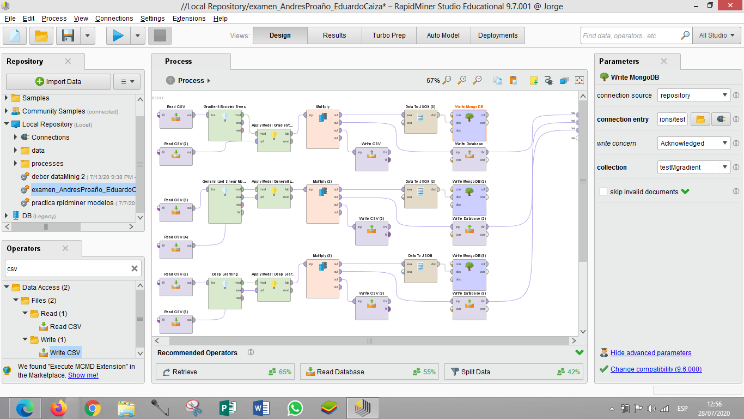
A continuacion tenemos el proceso para crear un archivo csv para lo cual simplemente escogemos el operador write y escogemos donde se quiere guardar el archivo



Ahora podemos ver que nuestros archivos csv fueron creados exitosamente como se puede ver en la imagen



Por ultimo tenemos una imagen que nos muetsra todas las conexiones realizadas ademas de un modelo final mediante el cual se pudo realizar las diferentes tareas asignadas, al momento de realizar la ejecucion del proceso se puede constatar que funciona correctamente



# CONLUCIONES

Como conclusión podemos decir que para realizar la predicción de los diferentes datos dependerá de cada modelo por lo que la precisión también dependerá del modelo ya que en algunos casos el uso del modelo ayudara a la precisión para que sea más alta

# Referencias

1. microsystem RAPIDMINER (online) Available: <https://www.microsystem.cl/plataforma/rapidminer/>

1. [↑](#footnote-ref-1)